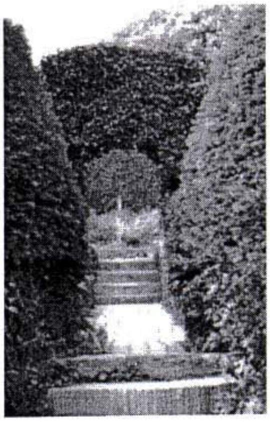
WEB DOCUMENT MANAGEMENT



Bibliography

This Web site is supported by a bibliography
Not everything in it is cited in the site.
You don't have to read it all.
It is a supporting resource.

For a brief overview of some of the issues covered by the course, take a look at the attached slideshow.

MODULE 1- Introduction	PAGE 1 Purpose	PAGE 2 Requirements	PAGE 3 Expectations	PAGE 4 Syllabus	PAGE 5 Assignments	PAGE 6 Calendar

MODULE 2 Web is What	PAGE 1 Web as Library	PAGE 2 Digital Libraries	PAGE 3 Info Architecture	PAGE 4 Equity

MODULE 3 Author	PAGE 1 Spam Indexing	PAGE 2 Mark Up	PAGE 3 MetaTags	PAGE 4 SGML/XML	PAGE 5 Dublin Core	PAGE 6 PICS	PAGE 7 Metametadata

MODULE 4 Characteristics	PAGE 1 URx	PAGE 2 Permanence	PAGE 3 Longevity

MODULE 5	PAGE 1	PAGE 2	PAGE 3	PAGE 4	PAGE 5	PAGE 6

Catalogs	Innovative Characteristics Concepts Relationships	MARC	GILS	Issues	OCLC	Pathfinder & Bookmarks

Module 6 Conclusions	Page 1 W3C MetaData	Page 2 Software	Page 3 R&D	Page 4 Concluding Remarks

The course calendar is marked with the module numbers and colors to indicate the approximate begin/end dates for each section.

Welcome to the Web Document Management Course. It is designed to be self-contained, asynchronous Web-based course offered by Dr. Wallace Koehler, MLIS Program, at the Valdosta State University.

The course is designed to familiarize graduate students and working librarians with the various approaches to and problems with bibliographic management of the Web strategies either at the author end or at the cataloger end.

The first question necessarily must be "can we do it", followed quickly by "should we do it." If we can do and if we should do it, what is it we should do? These define the purpose of this course.

As is the case with all introductory courses, to understand any component you must already understand all the others. This be particularly true of a course such as this. Fortunately for us in a Web-based course, all the components are present and can be referenced as needed.

I have organized the course as a series of modules. I have ordered the modules with a purpose. They are ordered with a mind to the information needed to understand what follows, assuming that the student has no prior knowledge of the subject matter. That will not always be the case, and if for example, you are an expert of RDF there is no need to "consume" that page.

With this in mind, we will not look at catalogs as such until we have explored how Web catalogs and indexes are created. This is divided into two groups: (1) *pre hoc* author cataloging and indexing (2) *post hoc* cataloging and indexing. Some indexing is automatic, some in manual. And so on. At times, I provide links to other site pages to support arguements and to avoid redundancy. You may move about the Web site/course as you wish.

A couple of things to remember. First, this is a survey course. In a scant semester it is not possible to go into full and technical detail about everything. The purpose is to offer a taste of what's going on and to give sufficient understanding of the dynamics of the field.

Second, this is complicated stuff. Take it slow. Explore the examples and explanations carefully. Go over them several times. When you can, generate your own mark up. I have tried to give examples that render the concepts as straightforward as possible.

Third, realize that this is copyrighted material. You may, however, download and/or print a copy of the Web site for your own educational purposes and retain it as a reference tool for yourself only.

Finally, please keep the following in mind as you work through this stuff. It was written in 1994, eons ago in Web-speak, it is valid then and it is valid now:

> "To oversimplify, the difference between a library and a pile of paper is how easy it is to find the particular items which interest you, especially when you don't know exactly what you're looking for in the first place. The tools currently available for sorting through the Internet haystack ... are much better than no tools at all, but they are still extremely crude when compared to the subtlety and precision of the average library catalog or journal index. It's worth asking why this is the case and what to do about it.
>
> "Librarians are at a disadvantage because, to generalize shamelessly about the profession as a whole, they can never quite understand the technology as well as the people who invent it. While many librarians are struggling valiantly to keep up with runaway technology, they are continually in a position of reacting to rather than originating change.
>
> "Computer people, for their part, have some handicaps as well. One handicap is a lack of understanding of library science. Computer people are continually trying to reinvent concepts which librarians have been honing for decades."
>
> Prentiss Riddle, 1994, revised 1996. *Library culture, computer culture, and the Internet haystack*
> Available: http://is.rice.edu/~riddle/dl94.shtml
>
> Another valuable resource. A slide series by Isidro Aguillo, *Evaluación de Recursos Web*, Centro de Información y Documentación Científica (CINDOC) del Consejo Superior de Investigaciones Científicas (CSIC)
> - in Spanish. Copyright 2000 Isidro Aguillo All Rights Reserved -- Used Here by Permission

---

In 1991 Louise Addis became the first librarian to appreciate the power of the Web as an information distribution tool. She is the first Internet cataloger [1, 2]. We have come a long way since Librarian Addis began serving up physics documents. The Web has gone from a few nodes to uncountable ones, it has gone from a few pages to more than a billion in the scant ten years since Addis began her Web library. This past ten years has witnessed the explosion of an information source from CERN to a certain impact on the lives of all information professional everywhere.

This course is an attempt to describe the strategies we now take to harness the power of the Web and to bring some kind of order to that incredible chaos that began as ARPANET in the 1960s, became the WWW in the 1990s, and will inevitably continue its growth. That continued growth will serve to challenge us all.

A new field is in the process of being defined -- Information Architecture. This course addresses "library parts" of the arena.
You will not be a fully qualified information architect when you finish the course, but you will begin to see what skills are needed to become one.

Web libraries proliferate

Web libraries can be defined as Internet resident resources that point to, collect, categorize, and/or catalog other Web resources. These can be distinguished from "libraries on the Web," like the US Library of Congress' National Digital Library Initiative, Project Gutenburg, or the many Internet interfaces to "brick and mortar" library collections. Libraries on the Web use the Internet as an access tool and as a pointer to their "traditional" resources, Web libraries not only use the Web as an access tool but also as the source of their collections.

Web libraries range in size, scope, quality, and depth from rudimentary jumppages linking to pages with similar content to sophisticated, monitored, and specialized collections like the Social Science Information Gateway or the Virtual Library collections. Library selection and de-selection may be performed by the Web author alone, by teams of subject specialists and librarians, or by automated procedures. The smaller libraries typically organize material by subject, author, or some other criterion, and provide links to the collected document. The larger libraries often provide site maps as well as limited area search engines to facilitate access and retrieval functions.

Several approaches have been taken to catalog the Internet. Most conclude that the Web is very elusive and difficult to catalog in large part because it is dynamic. There are essentially three approaches to Web cataloging.

- The first is to treat Web content as static. One variant is to remove non-responsive links. Many online libraries and Web resource pages test their collections for viability. A second variant is to little or no viability testing once published or created.
- The second is to archive a static collection and to catalog that collection. A variety of solutions have been offered, and these include the archive proposed by Kahle (1997) and URI, URN, and URC concepts.
- The third is to recognize the dynamic nature of the Web and to provide dynamic bibliographic representation of document content and location.

It has been clearly demonstrated in practice and in the literature that Web resources can and have been cataloged or indexed according to established library practice (Bates 1998). Yahoo!, an alphabetico-classed system, is a variation on colon classification and chain indexing first developed by S.R. Ranganathan in the early 1930s. Colon and faceted classification systems have the value of challenging universal classification, like the Dewey, Library of Congress, or BBK, and may find a further place in mapping the heterogeneous Web environment (see Star 1998).

Information retrieval systems equally or more sophisticated than Yahoo! have been developed and distributed on the Web. These include the Social Science Information Gateway (SOSIG) and the Scout Report Signpost. SOSIG is cataloged utilizing a metadata template and controlled vocabulary. A thesaurus supports its controlled vocabulary. The Scout Report is proof-of-concept research to demonstrate the utility of the application of standard library classification to the Web. It utilizes both Library of Congress classification and Subject Headings to supports its catalog (Glassel and Wells 1998). NetFirst, an outgrowth of InterCat, is one of the databases that make up OCLC's FirstSearch. NetFirst is a catalog of Web documents, classified according to established metadata, including Dewey Decimal Classification.

To paraphrase the old saw, there is a flaw in each of the ointments. The World Wide Web will not sit still. Not only is new material being constantly added, just as it is in the analog world; what already "is" is dynamic and changing. This has led some to consider whether the Web can ever be meaningfully cataloged (Oder 1998) or to despair that it cannot (Ardito 1998). Still others have developed guidelines for Web cataloging (OCLC nd; Olsen 1997).

Recognizing differences between the Web and other media, still others have taken an intermediate approach. Many Web libraries periodically sweep their collections and weed "dead links." Others, like the Scholarly Societies Project at the University of Waterloo (1999) select for inclusion only those URLs deemed likely to be "stable." Reporting their experiences in developing a Web catalog that quickly dated, McDonnell, Koehler, and Carroll (1999) argue the need for general cataloging of Web sites and against detailed cataloging of Web pages. Because of the ephemeral nature of Web material and the relative durability of Web sites and Web pages located on directory structures closest to the server address, they recommend that individual catalog entries be general and describe overall content vectors rather than the content of specific Web pages. In a similar vein, Tennant (1998) recommends "digital bibliography" over digital cataloging.

Reliance on robot indexed search engines is no answer either. If catalogs have a clear advantage over robot indexed search engines, it is because the former are often constructed based on a controlled vocabulary while the latter are not. It is precisely on this point that author prescribed indexing like Dublin Core, html metatags, and other document marking strategies are weakest.

It has also been clearly demonstrated in the literature that general search engines do not begin to cover the full extent of the Web (e.g. Tomaivolo and Packer 1996, Brake 1997, Koehler 1997, Lawrence and Giles 1998). Moreover, those same engines also demonstrate index timeliness erosion, they too become "stale" (Koehler 1998).

Any catalog that provides bibliographic access to "nothing" or to content-shifted documents is worse than no catalog at all.

Purpose,Expectations, and Requirements

LIS 5990 is designed to address means and methods to manage Web documents in the library and information science context. This includes a wide range of skills and techniques. We will

- Develop selection criteria for Web library creation
  - Can the WWW be managed
  - What are the quality issues
  - Is the WWW different
- Explore existing and proposed techniques for Web page management, indexing, and cataloging
  - Use of search engines as indexing applications
  - Metadata approaches to Web management
  - Creating and maintaining a Web "library"

- Archiving issues

We will explore search engines as "indexing tools." We will consider, examine, and apply various author-based and indexer-applied systems and technologies. We will consider problems unique to the Web environment and look to metadata applications that may provide some control for those elements.

This is a Web based course. All instruction, assignments, discussion, and other communication are designed for the electronic environment. There are necessarily certain assumptions that follow. All students must be conversant with a minimum set of electronic skills. These include use of email, discussion and chat groups, ftp, email attachments, etc. Students must also have or have available Internet technology and connectivity.

Be aware that the bibliographic management of the Web is a new art. But also remember that it has been the role of librarians since time immemorial to store, categorize, manage, distribute, and organize the information legacy of our societies and cultures. If there are any surprises arising from the new approaches, it is not how complex they are, but how new they are not. In fact, as you will gather throughout this course, I believe that as an information medium, the WWW is less advanced in concept than what we have "traditionally" employed. In some ways it is closer to the oral tradition than to the written.

....

SYLLABUS ....

..Contact

Dr. Wallace Koehler
MLIS Program
Valdosta State University
Valdosta, GA 31698
V: 229 245 3732 F: 229 259 5055
wkoehler@valdosta.edu

General Context

Is the WWW a library? Or is it a source of materials to be incorporated into libraries. As always there is a difference of opinion.

[1] Tim Berners-Lee with Mark Fischetti. Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor. NY: HarperSanFransisco, 1999, pp. 45-6.

[2] "FM Interviews Louise Addis." *First Monday*, 5, 5 (May 2000) http://www.firstmonday.org/issues/issue5_5/addis/index.html

BOBBY v3.1 APPROVED